DOCUMENT RESUME

ED 236 177 TM 830 681

AUTHOR Sax, Gilbert; Reiter, Pauline B.
TITLE Reliability and Validity of Two-Option Multiple-Choice and Comparably Written True-False Items.
PUB DATE [80]
NOTE 12p.
PUB TYPE Reports - Research/Technical (143)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Analysis of Variance; Correlation; Higher Education; *Item Analysis; Multiple Choice Tests; Objective Tests; *Test Format; *Test Items; *Test Reliability; *Test Validity

ABSTRACT

Despite the popularity of both multiple-choice (MC) and true-false (TF) items, most investigations comparing the two formats have done so to determine the optimum number of choices to be given to students within a given time period. The purpose of this investigation was to compare the reliabilities and the validities of both formats when the items were identical except that the MC format presented examinees with a correct and incorrect option while the TF format included one of these two options in the stem. On the TF forms, students responded to a TRUE or to a FALSE option. Items in both formats were further analyzed by developing a MC and TF form having high point biserial correlations; another set of items used distractors chosen at random (all items had been administered previously). A one-way Analysis of Variance (ANOVA) was significant at the .01 level with the MC items yielding the higher means. No differences were found among the Kuder-Richardson reliability coefficients or among the validity coefficients (measured by correlating test scores with overall grade point average). Additional subjects have been tested, and data are being analyzed with a 2X2 factorial ANOVA (MC-TF and two levels of discrimination). (Author/PN)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☒ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
G. Sax
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

ED236177

TM 830 681

Reliability and Validity of Two-Option Multiple-Choice and Comparably Written True-False Items

Gilbert Sax and Pauline B. Reiter

University of Washington

True-false (TF) items have been classified as both multiple-choice (MC) and as alternate response items. Mehrens and Lehmann (1973), for example, described TF items as being two option MC items that call for a true-false, right-wrong, or yes-no response where only one of the propositions is given. Thorndike and Hagen (1969), however, considered TF items to be alternate-choice items; to qualify as a MC item, they required a minimum of three alternatives.

The greater versatility, the decreased probability of guessing correctly, and the ability to reduce ambiguity have contributed to the popularity of MC items (Gronlund, 1976; Noll and Scannell, 1972; and Sax, 1980). TF items, despite their apparent simplicity and efficiency (Ebel, 1965; 1972), have been criticized as being limited in application because of their dependence on absolute truth or falsity (Sax, 1980); they are often ambiguous to students, unclear, and they are affected excessively by high chance level scores that tend to reduce reliability (Gronlund, 1976; Mehrens and Lehmann, 1973). Ebel (1972) has defended the TF item by maintaining that their faults lay with the item writers and not with the TF items themselves.

In studies conducted by Frisbie and Ebel (1972) and by Oosterhof and Glasnapp (1974), their findings favored the MC test format ($p<.01$). These investigators did not, however, compare two-option

MC items with comparably written TF items. Instead, Frisbie and Ebel asked teachers to judge the most plausible incorrect option in one phase of their study; in another phase, they constructed a false item by selecting distracters with the largest upper-lower differences. Thus, TF items were developed either by teacher subjectivity or by empirical procedures that depended originally on the quality of the MC distracters. KR[20] reliabilities were then adjusted by the Spearman-Brown formula to compensate for inequalities in response time. Although the MC and TF items were designed to measure the same objectives, items differed in wording, quality of distracters, and time required to complete the tests.

Oosterhof and Glasnapp (1974), Straton and Catts (1980), and Irvin, Halpern, and Handman (1980) were also concerned with the reliabilities of tests having differing numbers of options. In each instance, however, the TF and MC items differed in wording. Additionally, Irvin, Halpern, and Handman administered their tests orally to retarded high school students--a condition likely to penalize students given MC items. Oosterhof and Glasnapp compared TF with 4-option MC items by administering both forms to all students. An illustration provided by the authors clearly demonstrates that the MC and TF items were not written comparably--a condition made necessary by the administration of all forms to all students. Items were modified to reduce the probability that they would be identified.

The purpose of the present investigation is to compare the reliabilities and concurrent validity coefficients of TF and

identically written 2-option MC tests. The hypothesis is that reliability and concurrent validity will be significantly larger for 2-option MC tests than for TF tests since MC items provide the examinee with a specific comparison while TF items require examinees to respond to independent statements of truth or falsity.

PROCEDURES

Subjects. Data were obtained from 62 upper division and graduate students who were enrolled in two separate introductory classes in statistics at the University of Washington. The two classes were taught three years apart by the same instructor who used the same text, class notes, and examinations throughout the quarter. Within each class, students received either the TF or the MC form of the final examination in a completely random fashion.

Instruments. A 46-item pool was assembled that consisted of 4- option MC items that had been administered at various times in previous years. The best incorrect answer for each item was determined by selecting the most discriminating distracter (i.e., the highest point-biserial r). The TF answers on the 46-item TF test corresponded to the correct and incorrect options on the MC test. The false responses contained the exact wording of the distracters on the MC test, and the true responses contained the exact wording of the correct responses on the MC items.

Example of a MC Item:

The true limits of 12.4 pounds are

A) 12.35 to 12.45 (option A is the correct response)

B) 12.3 to 12.5

Example of a Comparably Written TF Item

The true limits of 12.4 pounds are 12.35 to 12.45

A) True (option A is the correct response)

B) False

In a second phase of this study, distracters (incorrect responses) were chosen at random, and the point-biserial correlations were disregarded. Data were analyzed separately for each of the two parts of the study.

METHODS

Students within each class were assigned randomly to one of four test forms: 1) TF format with the options containing the distracter with the highest point-biserial r [TF best]; 2) TF format with an option selected at random [TF random]; 3) MC with the best distracter included as a foil along with the correct answer [MC best]; and 4) MC with a distracter selected at random along with the correct response [MC random].

Four forms of the test were prepared that were identical with the exception of item format and the use of different options. Within each of the classes that participated in this study, the four forms were shuffled thoroughly and distributed randomly to each student. Students were allotted one hour and thirty minutes to complete their tests. Because these time limits were generous, all students were easily capable of completing the test within the allotted time limits. Students were unaware that they were part of a study. It was for that and for other ethical reasons that at

the end of the examination, students were debriefed. For marking purposes, the same proportions of students were given equivalent grades. Students were asked to indicate their overall grade-point average on their answer sheets. The few students who did not comply with that request were eliminated from the study if they so requested; otherwise, students gave permission in writing for us to obtain those data from the registrar's office. As a validity check, a random selection of 15 estimates of GPA was compared with official transcripts. Differences between estimates of GPA and data provided by the registrar were both statistically insignificant and of little practical importance. Slight overestimates of GPA were provided by students with MC and with TF examinations. For the MC group, the mean overestimate was .04 with s=.13, N=8; for the TF group, the mean overestimate was .03 with s=.14, N=7; t=.14, ns.

Because of the ample time limits, corrections for test length were judged to be unnecessary especially since the amount of reading on the four forms was virtually identical.

RESULTS

Table 1 presents the summary statistics for each of the four forms and for the combined TF and MC formats.

Table 1 here

A one-way ANOVA was used to test the null hypothesis that the means of the 4 item formats were drawn from the same population. With 3 and 56 degrees of freedom, the F-ratio was 9.14 with $p<.01$. A Scheffe' test demonstrated significant differences ($p<.01$) between

the following paired comparisons: MC best > TF best; MC total > TF total. In both instances, then, the highest means were obtained by the MC tests.

Feldt's test was used to evaluate the differences between the Kuder-Richardson reliability coefficients. Although all pairs were examined, none was statistically significant. Nonetheless, the differences in reliability between the TF-best and the TF-random is equivalent to a fourfold increase in the number of items as estimated by the Spearman-Brown formula. Using the same reasoning and procedure, the MC-total reliability is equivalent to what might be expected from the TF-total if the number of items on the latter format were doubled.

Table 2 presents data pertaining to the concurrent validity of

Table 2 here

each of the four forms of test formats with overall GPA as the dependent variable. Although none of the formats proved to be superior with regard to validity (all t-tests for correlations between test format and GPA were not significant statistically) the validity of the TF-random format was the highest while the TF-best resulted in the lowest coefficient.

DISCUSSION

The intent of this study was to compare four differing test formats--TF-best, TF-random, MC-best, and MC-random-- as to their reliabilities and validity coefficients. The highest Kuder-Richardson reliabilities were derived from the two MC tests. The

lowest reliability was computed from the TF-best items. The most plausible explanation for these findings concerns the means of the various item formats. Relatively easy tests tend to yield high reliability coefficients. Because the TF-best form consisted of items with high point-biserial coefficients, they tended to be difficult; the random form, in contrast, being easier, also tended to yield high reliability. In each instance, whether the items were in MC or in TF formats, the random form yielded higher coefficients. Of some interest is the finding that MC examinations of the type administered for this study tended to yield almost identical means as well as Kuder-Richardson reliabilities. When combined into single forms of MC and TF items, the MC items as a whole tended to be easier than did the TF items, and, as a result, the MC form was the more reliable test.

More difficult to explain are the validity coefficients. The largest validity coefficient was produced by the TF-random items which also had the smallest standard deviation in the criterion measure, GPA. The TF-best items, which had the lowest reliability, also had the lowest validity even though the standard deviation of the criterion measure was among the highest of the various GPAs. Although the differences in the standard deviations of the six forms are small, they correlate highly with the validity coefficients ($r=.85$). In part, of course, the small Ns within each group make it difficult to be confident about any explanation; still, classrooms of 31 are common, and when those data are examined it seems reasonable to conclude that the MC format is superior to

the TF format in both reliability and validity.

This study should only be considered as a beginning of a larger investigation into the relative advantages and disadvantages of different types of TF and MC items. As such, it represents a progress report and not a completed investigation. Additional subjects have already been tested on the four forms of test items at the end of winter quarter, 1983. Time has not permitted us to analyze these data as yet, but another 10 students within each group should provide for more reliable information.

Table 1:

Means, Variances, Reliabilities, and Validity Data

Obtained on MC and TF Tests

Item Format	N	$\bar{X}$	S^2	S	KR_{20}	$SE_{meas.}$
TF-Best	15	27.67	12.11	3.48	.24	3.04
TF- Random	16	30.25	16.81	4.10	.54	2.80
MC-Best	16	34.44	18.84	4.34	.64	2.61
MC-Random	15	34.13	18.40	4.29	.67	2.45
TF-Total	31	29.00	16.16	4.02	.45	2.99
MC-Total	31	34.29	18.66	4.32	.63	2.61

Table 2:

Validity Data

Item Format	$\bar{X}$ GPA's	S^2 GPA's	S GPS'S	r_{xy}
TF-Best	3.37	.27	.52	.10
TF-Random	3.49	.11	.33	.49
MC-Best	3.43	.34	.58	.45
MC-Random	3.42	.17	.41	.35
TF-Total	3.43	.19	.43	.29
MC-Total	3.43	.25	.498	.41

References

Costin, Frank. "The Optimal Number of Alternatives in Multiple-Choice Achievement Tests: Some Empirical Evidence for a Mathematical Proof." Educational and Psychological Measurement, 30, 1970, 353-358.

Ebel, Robert L. Measuring Educational Achievement. Englewood Cliffs, N.J.,: Prentice-Hall, Inc., 1965.

Ebel, Robert L. Essentials of Educational Measurement.Englewood Cliffs, N.J.: Prentice-Hall, Inc., 1972.

Frisbie, David A. and Ebel, Robert L. "Comparative Reliabilities and Validities of True-False and Multiple Choice Tests." Chicago: AERA, 1972.

Gronlund, Norman E. Measurement and Evaluation in Teaching. New York: Macmillan Publishing Co., 1976.

Irvin, Larry K.,Halpern, Andrew S., and Handman,Janet T. "Assessment of Retarded Student Achgievement with Standardized True/False and Multiple-Choice Tests." Journal of Educational Measurement, 17, 1980, 51-58.

Mehrens, William A. and Lehmann, Irvin J. Measurement and Evaluation in Education and Psychology. New York: Holt, Rinehart, and Winston, Inc., 1973.

Noll, Victor H. and Scannell, Dale P. Introduction to Educational Measurement. Boston: Houghton Mifflin Co., 1972.

Oosterhof, Albert C. and Glasnapp, Douglas R. "Comparative Reliabilities and Difficulties of the Multiple-Choice and True-False Formats," Journal of Experimental Education, 42, 1974,62-64.

Sax, Gilbert. Principles of Educational and Psychological Measurement and Evaluation. Wadsworth Publishing Co., 1980.

Straton, Ralph G. and Catts, Ralph M. "A Comparison of Two, Three, and Four-Choice Item Tests Given a Fixed Total Number of Choices," Educational and Psychological Measurement, 1980, 357-365.

Thorndike, Robert L. and Hagen, Elizabeth. Measurement and Evaluation in Psychology and Education. New York: John Wiley, 1969.

Reliability and Validity of Two-Option Multiple-Choice and Comparably Written True-False Items

ABSTRACT

Despite the popularity of both multiple-choice (MC) and true-false (TF) items, most investigations comparing the two formats have done so to determine the optimum number of choices to be given to students within a given time period. In these studies, little attention was paid to the lack of comparability of the items themselves. The purpose of this investigation was to compare the reliabilities and validities of both formats when the items were identical except that the MC format presented examinees with a correct and incorrect option while the TF format included one of these two options in the stem. On the TF forms, students responded to a TRUE or to a FALSE option.

Items in both formats were further analyzed by developing a MC and TF form having high point biserial correlations; another set of items used distracters chosen at random (all items had been administered previously). A one-way ANOVA (dfs = 3,58) was significant at the .01 level with the MC items yielding the higher means. No differences were found among the KR_{20} coefficients or among the validity coefficients (measured by correlating test scores with overall GPA). Additional subjects have been tested, and data are being analyzed with a 2X2 factorial ANOVA (MC-TF and two levels of discrimination).